

**Review Exercises Chapter 10**

- #2 (a) The individuals who scored 115 at age 18 scored exactly 1  $SD_{18}$  above the mean back then. From our regression formula we know that # of  $SD_{35} = r \times (\# \text{ of } SD_{18}) = 0.80 \times 1 = 0.80$ . Thus, the average score at age 35 for all individuals who scored 115 at age 18 is 0.8 standard units above the overall average score at age 35. To see how many IQ-points at age 35 this is, we multiply the # of  $SD_{35} \times SD_{35} = 0.8 \times 15 = 12$ . Now we add 12 to the mean IQ-Score of age 35. That is,  $\mu_{35} + 12 = 100 + 12 = 112$ .  
Combining the steps above in one single equation yields

$$\mu_{35} + \left( \frac{115 - \mu_{18}}{SD_{18}} \times r \right) \times SD_{35} = 100 + (1 \times 0.80) \times 15 = 100 + 12 = 112.$$

- (b) The computation here is exactly the same as in (a), since our estimate of the average score at age 35 is also our best guess given the information. Thus, we predict that her score will be 112 at age 35.
- #4 (a) Based on the information given, our best guess of the wife's education level is

$$\begin{aligned} \mu_{wives} + \left( \frac{18 - 12}{3} \times r \right) \times SD_{wives} &= 12 + (2 \times 0.5) \times 3 \\ &= 12 + 3 = 15 \text{ years} \end{aligned}$$

- (b) Here we know that the wife has an educational level 1 SD above the wives' mean education level, so our best guess of her husband's years of schooling is

$$\mu_{husbands} + (1 \times r) \times SD_{husbands} = 12 + (1 \times 0.5) \times 3 = 13.5$$

- (c) This is because two regression lines can be drawn across the scatter diagram, one for predicting the wife's education level from her husband's, and one for predicting the husband's education level from his wife's. (See figure 9, page 175.) From the data we are given, we can calculate each of these regression lines. They are:

$$\text{husband's years of education} = 6 + (.5 \times \text{wife's years of education})$$

$$\text{wife's years of education} = 6 + (.5 \times \text{husband's years of education})$$

The reason we cannot simply perform algebra to solve for the wife's years of education from the first equation is that the husband's years of education and the wife's years of education are not perfectly correlated.

If we look at the strip of the scatter plot containing the most well-educated men, the average education level for their wives will be lower than the well-educated husbands. Similarly, if we look at the strip of the scatter plot containing the most well-educated women, the average education level for their husbands will be lower than the well-educated wives. Thus, we would predict that a well-educated man would marry a woman less educated than him, and a well-educated woman would marry a man less educated than her.

- #5 (a) False. The correlation coefficient of 0.6 tells us the degree to which the observations cluster about the regression line; this is *not* the slope of the regression line. We are not given enough information to calculate the regression line and so cannot make predictions of an athlete's lifting ability.
- (b) False. Because we do not know the standard deviations or the means for each variable, we cannot estimate the effect of an athlete's weight increase on her weight lifting ability.
- (c) True. The correlation coefficient of 0.6 is positive, which means that the weights athletes can lift and the weights of the athletes have a positive relationship. As one increases, so does the other. Therefore, the more an athlete can lift, the more he weighs on average.
- (d) True. The correlation coefficient is positive, which means that the variables have a positive relationship. Therefore, the more an athlete weighs, the more he can lift on average.
- (e) False. Correlation measures association, not causation.

#8 The study suggests that patients are more relaxed on the second reading. If the regression effect was at work, we would expect the average blood pressure to be about the same for both the first and second reading. In this case, the average blood pressure falls by about 10mm from the first to the second reading.

- #9 (a) Clearly, this student is in the left-tail of the distribution. In order to predict his/her likely rank on the final exam, we have to find the z-score for his/her midterm first. We know that at the student's percentile rank 5% of all observations are located in the distribution's left tail. Assuming that the exam scores are normally distributed, we thus look up the z-score for an area of 90% in the normal table of the book (90% because by symmetry there are also 5% in the upper tail), which yields a z-score of approximately 1.65. However, as we are interested in the left tail,  $z = -1.65$ . (In other words, the student scored 1.65 SDs below average on the midterm.) The regression method now predicts that the student will be  $r \times (\# \text{ of } SD_{\text{midterm}}) = (0.50) \times (-1.65) = -0.825$  standard deviations away from the mean on the final exam. Looking at the normal table in the book, this corresponds to an area of about 59%, i.e., 59% of all observation fall between 0.85 and  $-0.85$ . Thus, about 20.5% of all observations are in each tail ( $\frac{100-59}{2} = 20.5$ ). Therefore, we predict that the student whose percentile rank on the midterm was 5% will do considerably better on the final exam and will on average obtain a score corresponding to the 20.5<sup>th</sup> percentile.
- (b) The approach here is the same as in (a). An 80<sup>th</sup> percentile rank gives us a  $z = 0.85$  (60% of all observation fall between  $\pm 0.85$ ). We then predict a score  $(0.50) \times (0.85) = 0.425$  SDs above the mean of the final, which translates to about 33.5% of all observations between 0.425 and  $-0.425$ . Thus, this student's percentile rank on the final exam is approximately  $\frac{33.5\%}{2} + 50\% = 66.75\%$ .
- (c) A student whose percentile rank on the midterm was 50% was of course 0 SDs away from the mean on the midterm. We thus predict no change for this student's percentile rank on the final exam, i.e., our prediction is that this student will be located at the 50<sup>th</sup> percentile on the final exam.
- (d) Our best guess of a student's performance on the final with unknown midterm performance, is the average student's final score. Since the scatter diagram is football-shaped, we know that the grades are normally distributed with a single mode. Thus, the mean will be equal to the median, leading us to predict that this student will be located at the 50<sup>th</sup> percentile rank on the final.

**Review Exercises Chapter 11**

#3 To answer this question, recall that the r.m.s. error of the regression line of  $y$  on  $x$  is given by  $\sqrt{1 - r^2} \times SD_y$ .

(a) In this case,  $y$  is height at 18, so we calculate

$$\sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.8^2} \times 2.5 = 1.5.$$

The r.m.s. error of the regression prediction is 1.5 inches.

(b) In this case,  $y$  is height at 6, so we calculate

$$\sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.8^2} \times 1.7 = 1.02.$$

The r.m.s. error of the regression prediction is 1.02 inches.

#4 (a) The question is asking you to determine the point value  $p$  such that the difference between a student's predicted score and actual score will be less than  $p$  for  $2/3$  of the students, and greater than  $p$  for the other  $1/3$ . In other words,  $2/3$  of observations fall within  $-p$  and  $p$  of the regression line. This should sound like a problem that requires  $Z$ -scores, like the questions from chapter 5. In particular, we want to use  $Z$ -scores to calculate the properties of the distribution of *prediction errors* (aka the residuals), the difference between the predicted and actual score for each student.

We don't need to perform any calculations to get the mean of the prediction errors: it must be 0, as described in section 3 of chapter 11 in the textbook. The standard deviation of the prediction errors—the average distance between the actual and predicted values—is just the r.m.s. error, as described in section 1 of chapter 11. Since  $r$  and  $SD_y$  are given, we can compute the r.m.s. error for the regression in this problem as

$$\sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.6^2} \times 15 = 12 \text{ points.}$$

Our first step, as in the problems in chapter 5, is to find the  $Z$  score such that about two-thirds (67%) of a standard normal curve is between  $-Z$  and  $Z$ . We can see from the  $Z$  table in the textbook that the appropriate  $Z$  score is somewhere between 0.95 and 1, so we'll use 0.975 as a guess. (It should make sense to you

that the  $Z$  score for 67% would be very close to 1, since we know from the empirical rule that 68% of the standard normal curve is between  $-1$  and  $1$ .)

The final step is to convert the  $Z$  score back into the units of interest. The  $Z$  score tells us that 67% of the prediction errors will be within 0.975 SDs of the mean. Therefore, in the original units, this means 67% of the prediction errors will be within  $Z \times \text{SD} = 0.975 \times 12 = 11.7$  points of 0. So our best guess out of the options given in the question is 12.

- (b) The students who scored 80 on the midterm are better than average. As a group they are expected to do better than average on the final exam – although there is a fair amount of spread as suggested by the correlation coefficient. Their average on the final can be estimated by the regression method: 80 is 1.2 SDs above average, so these students will score about  $r \times 1.2 = 0.6 \times 1.2 = 0.72$  SDs above average on the final exam. This is  $0.72 \times 15 = 10.8$  points above average. So their predicted score is  $55 + 10.8 = 65.8$ .
- (c) Students, who scored 80 on the midterm are a smaller and more homogenous group than the whole class. So the SD of their final scores will be less than 15. How much less? Since the diagram is football-shaped, the scatter around the regression-line is about the same in each vertical strip and is given by the r.m.s. error for the regression line. The SD of the prediction in (b) is therefore

$$\sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.6^2} \times 15 = 12 \text{ points.}$$

Thus, the prediction is likely to be off by 12 points or so.

#9 The data do not conclusively support the notion of a “sophomore slump.” What is most likely at work here is the regression effect, which would lead us to predict that unusually high observations (in this case, a batting average of .285) will regress towards the mean in another round of measurements, regardless of players being “distracted” by TV appearances and the like.

#10 No, the regression method was not used to make the predictions. This can be seen by comparing the actual 2005 prices with the predicted 2006 prices. The first, and probably most intuitive way to approach the problem, is to plot the two variables in a scatter diagram. If the

predictions were derived by the regression method, you should be able to connect the five points by a straight line. This is clearly not the case.

The second way to think about this problem exploits the linear property of the regression method. If the regression method was used, then an increase of 1 in the 2005 price should be associated with a change of  $x$  in the predicted 2006 price, and an increase of 2 in the 2005 price should be associated with a change of  $2x$  in the predicted 2006 price. Looking at the data we can easily see that this is not the case: an increase of 1 from stock D to stock E in the 2005 price column is associated with an increase of 8 in the 2006 prediction column. However, an increase of 2 from stock C to stock D in the 2005 price column is associated with a decrease of 1 in the 2006 prediction column.

#12 False. This question requires us to first predict what level of blood pressure the man under scrutiny is expected to have, given his level of education, and then to establish a range of values that can be said to be typical for men at his educational level. Only if the predicted value falls outside this “normal” range can we conclude that his blood pressure level is unusual.

First, predict the expected blood pressure for a man with 20 years of education. 20 years of education is  $\frac{7}{3}$  SDs above the average education level. Using the information provided in the question, we can now predict the man’s blood pressure level:

$$119 + r \times \frac{7}{3} \times SD_{blood\ pressure} = 119 + (-0.1) \times \frac{7}{3} \times 11 = 116.4\text{mm}$$

Next, we have to find the r.m.s. error for this prediction. This is

$$r.m.s.\ error = \sqrt{1 - (-0.1)^2} \times 11 = 10.9\text{mm}.$$

Now we are finally in a position to judge whether a blood pressure level of 118mm is a bit on the high side for the man with 20 years of education. Based on our forecast of 116.4mm and the fact that about 68% of the actual blood pressure level values are expected to be within a range of  $\pm 10.9$ mm around this value, 118mm does *not* appear to be on the high side. It seems to be a rather typical value.

## Review Exercises Chapter 12

- #1 This question is most easily answered by first computing the slope of the regression line. To do so, we can use the formula from page 204 in FPP:

$$\text{slope} = \frac{r \times SD_{final}}{SD_{midterm}} = \frac{0.60 \times 20}{10} = 1.2$$

Next, we can use this information to find the intercept. We know that the regression line has to go through the point of averages, which allows us to write

$$\begin{aligned} \text{mean}_{final} &= \text{intercept} + 1.2 \times \text{mean}_{midterm} = \text{intercept} + 84 \\ \Rightarrow \text{intercept} &= 55 - 84 = -29. \end{aligned}$$

All parts of the regression equation have now been identified. Thus, the regression equation is

$$\widehat{\text{final score}} = -29 + 1.2 \times \text{midterm score}$$

- #3 If someone puts on 20 pounds, we do *not* predict that he will get taller by 0.5 inches, since weight obviously does not *cause* changes in height. It is only associated with height. What the slope of the regression line means in this case is that we predict that all those who are 20 pounds heavier than this person will on average be about 0.5 inches taller.
- #5 False, the second investigator has not found the regression equation predicting husband's income from wife's income. You cannot simply re-arrange the terms in the regression formula, because the standard deviations of the husband's and wife's incomes differ.

You can confirm this by recalculating the slope for the regression to predict husband's income ( $x$ ) from wife's income ( $y$ ). The new slope is

$$r \times \frac{SD_x}{SD_y} = 0.25 \times \frac{39,000}{26,000} = 0.375,$$

which is not what the second investigator calculated.

- #7 Recall that the regression line must go through the "point of averages," where  $x = \bar{x}$  and  $y = \bar{y}$ . In this problem, the average number of pizzas consumed is  $\bar{x} = 4$  and the average number of beers consumed is  $\bar{y} = 4$ .

We can substitute these into the given regression equation and solve for the missing slope  $b$ :

$$\begin{aligned}\bar{y} &= b \times \bar{x} + 2 \\ 4 &= b \times 4 + 2 \\ b &= \frac{4 - 2}{4} = \frac{1}{2}.\end{aligned}$$

The slope is therefore  $\frac{1}{2}$ .

#9 What is described here is an approximation to what is done in the regression method. Thus, the slope of that line is an approximation to the slope of the regression line, so we can use our formula for the regression slope, which is

$$b = \frac{r \times SD_{IQ}}{SD_{Income}} = \frac{0.50 \times 15}{\$45,000} = \frac{1}{6,000}$$